Capstone Project

Credit Card Default Prediction

Anson Sibi

Data Science Trainee , Almabetter

Problem Statement

❏ This project is aimed at predicting the case of customers default payments in Taiwan. From the perspective of risk management, the result of predictive accuracy of the estimated probability of default will be more valuable than the binary result of classification - credible or not credible clients. We can use the K-S chart to evaluate which customers will default on their credit card payments

Feature Summary

- ➢ X1 -Amount of credit(includes individual as well as family credit)
- ➢ X2 -Gender
- ➢ X3 -Education
- ➢ X4 -Marital Status
- ➢ X5 -Age
- ➢ X6 to X11 -History of past payments from April to September
- ➢ X12 to X17 -Amount of bill statement from April to September
- ➢ X18 to X23 -Amount of previous payment from April to September
- ➢ Y -Default payment

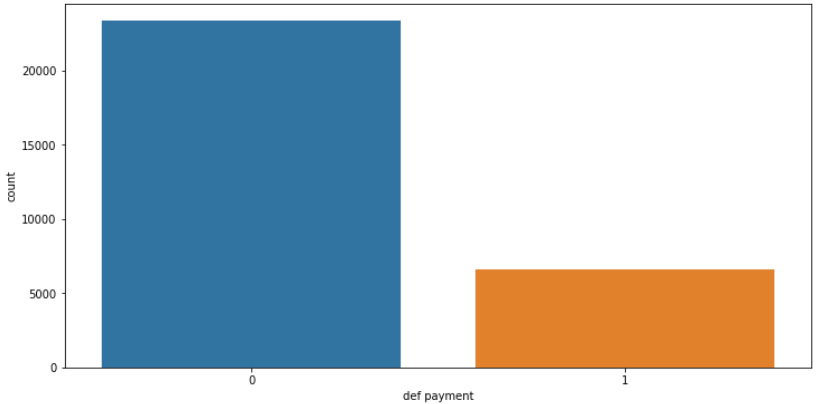
Data Exploration

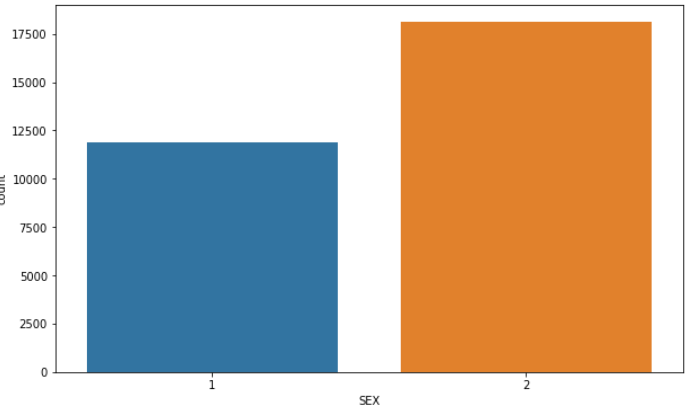
- ➢ This Dataset is from Taiwan.
- ➢ In our data set there are 30000 rows, 26 columns
- ➢ There are No Missing Values present
- ➢ There are No Duplicate values present
- ➢ There are No null values.
- ➢ And finally we have ‘def payment' variable which we need to predict for new observations
- ➢ 9 Categorical variables present.

Exploratory Data Analysis

Univariate Analysis

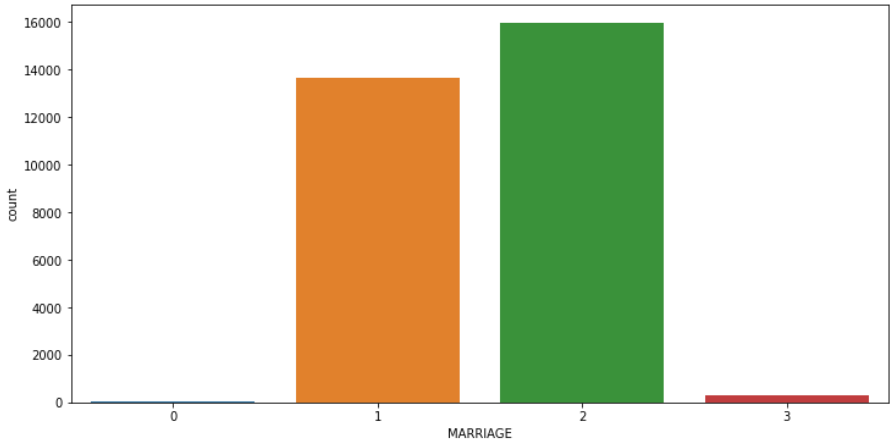
The number of Non-defaulters is greater than the number of defaulters. There is an imbalance between the classes.

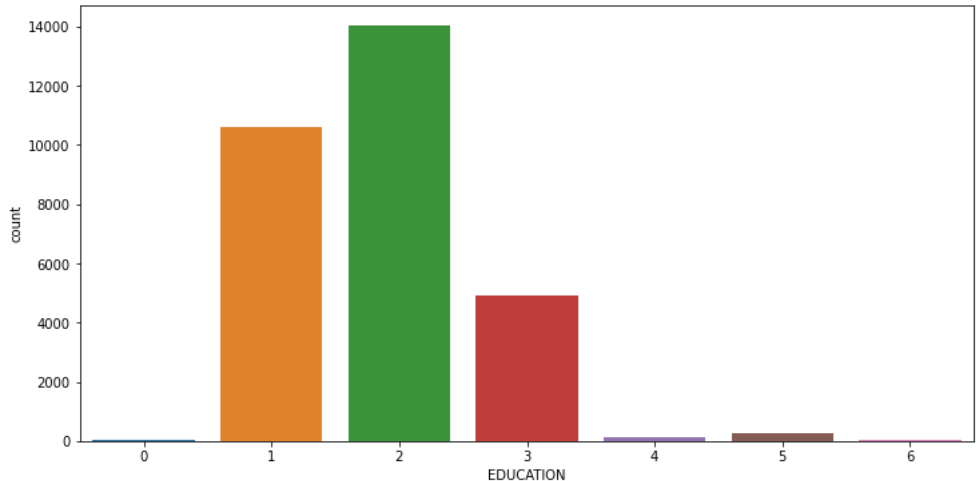




Female credit holders are more than male credit holders

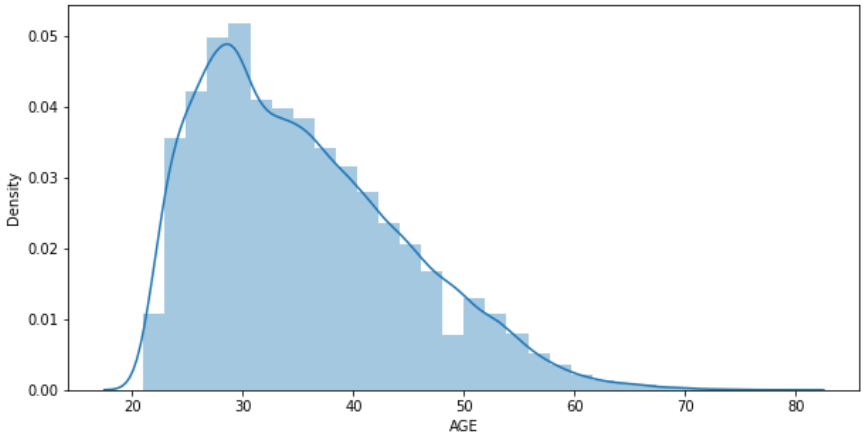
Most of the customers are either Married(1) or single(2) from the data.

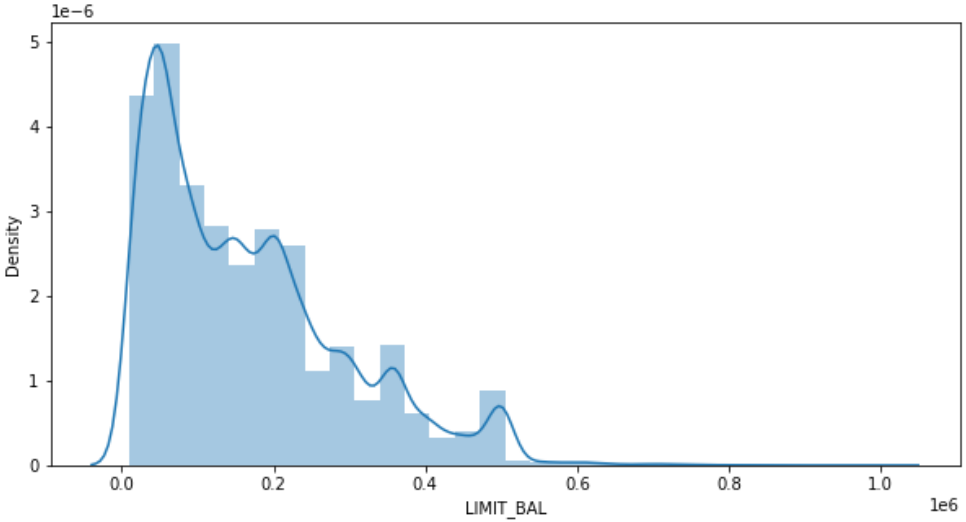




The majority of the customers have attained a university degree(2) followed by Graduate school(1) and high school(3).

Numerical Variables

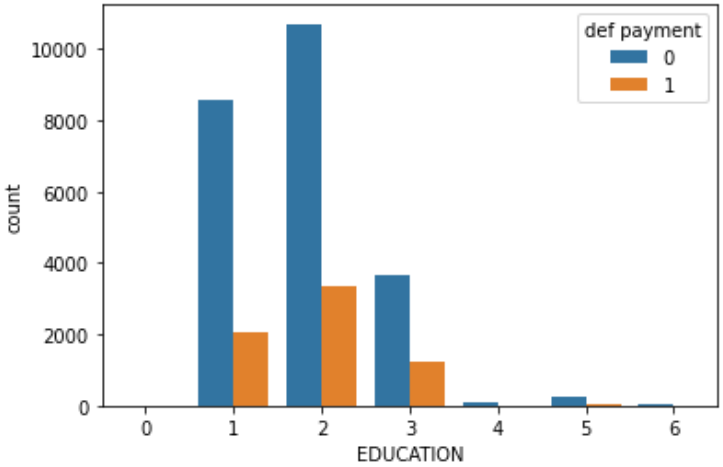




- Most people in our dataset have between 25 and 40 years old. There is also an impression that around that age the chan ce of default is a little lower.

- Most customers have 200k or less of credit limit. And it seems that we will find a higher concentration of customers in default on that range.

Bivariate Analysis

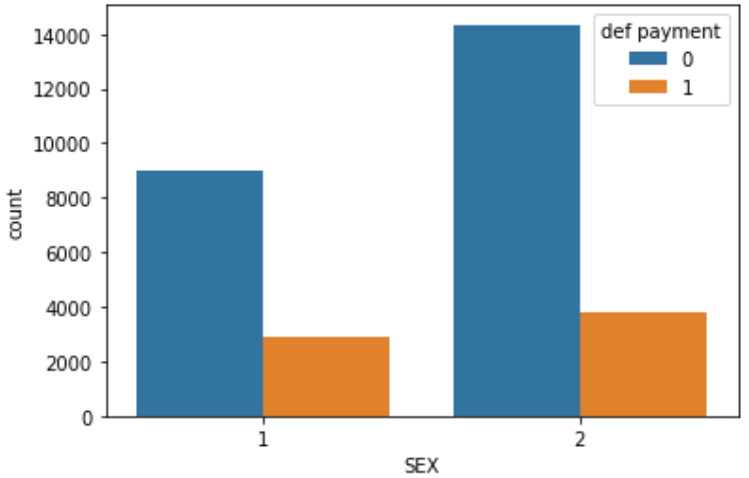


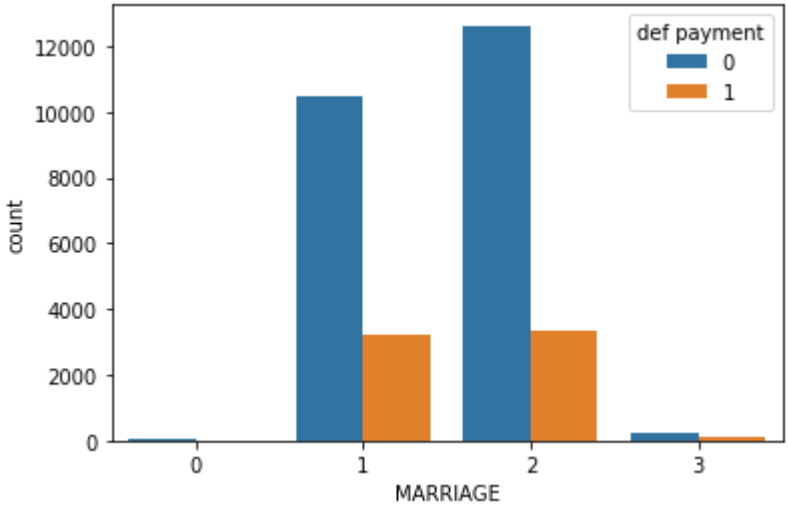
EDUCATION VS DEF PAYMENT

The predominant level of education in our dataset is (1 = graduate school; 2 = university; 3 = high school; 4 = others ;and (0,5,6)=Unknown)
Considering only the first three levels, it seems that a higher education translates to a lower chance of default.

SEX VS DEF PAYMENT

There are more women than men in our dataset and, apparently, men have a slightly higher chance of default.

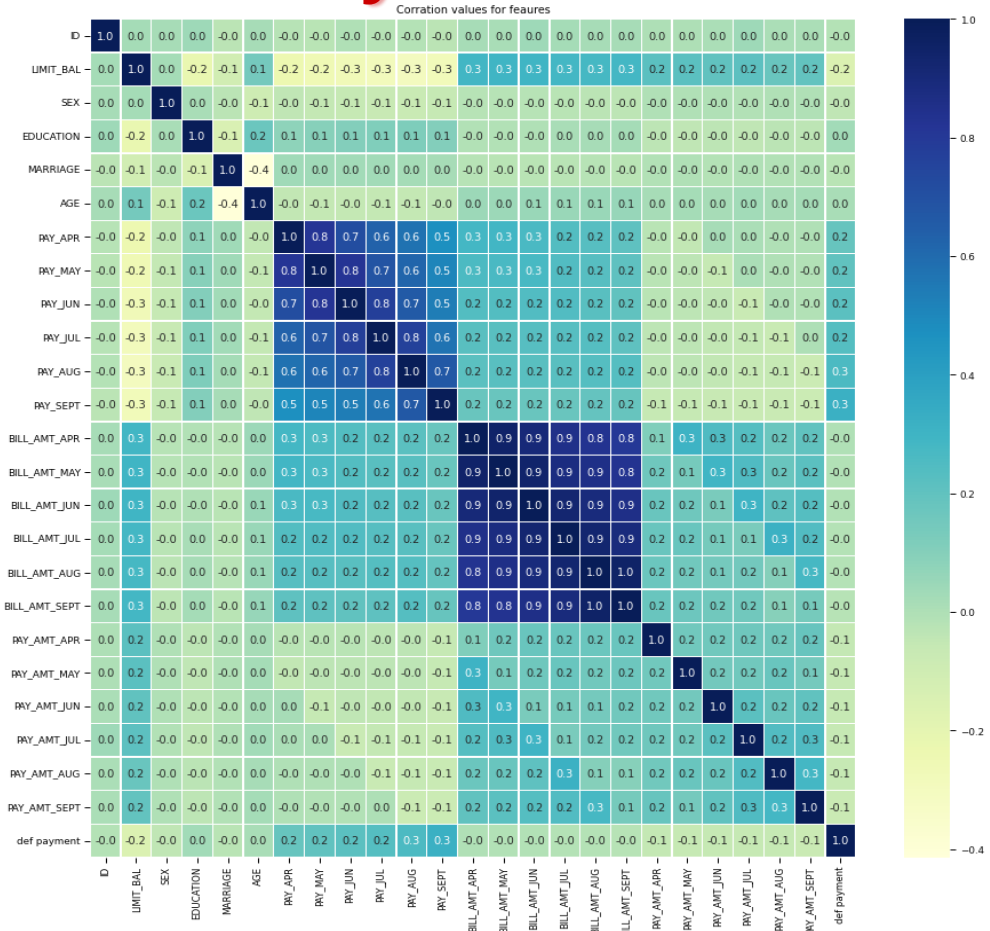




MARRIAGE VS DEF PAYMENT

Married people have higher number of defaulters as compared to single and others from the barplot given.

Multivariate Analysis



Machine Learning Model Insights

Comparing Model Performance Before Hyperparameter Tuning

	Model	Train Accuracy	Test Accuracy	Precision	Recall	F1 Score	ROC AUC score
0	Logistic Regression	0.753768	0.750909	0.789361	0.684128	0.732987	0.750876
1	DecisionTreeClassifier	1.000000	0.739710	0.732318	0.755210	0.743588	0.739718
2	RandomForestClassifier	1.000000	0.839789	0.860061	0.811447	0.835047	0.839775
3	XGB Classifier	0.774619	0.769955	0.809563	0.705681	0.754061	0.769923

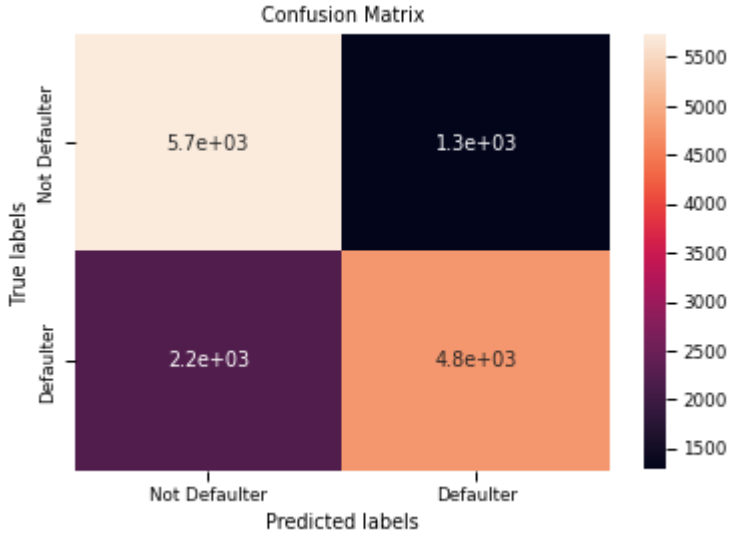
The Recall and Precision of the Decision Tree Classifier and Random Forest Classifier are the highest.
But there is a big difference between train and test accuracy which suggest that the model is overfitting.

Logistic Regression

	Model Name	Accuracy Score	Precision Score	Recall Score	f1 score	roc_auc_score
)	Logistic Regression	0.750909	0.789361	0.684128	0.732987	0.750876

The first tried model is Logistic Regression. The model is so simple and didn't perform well.

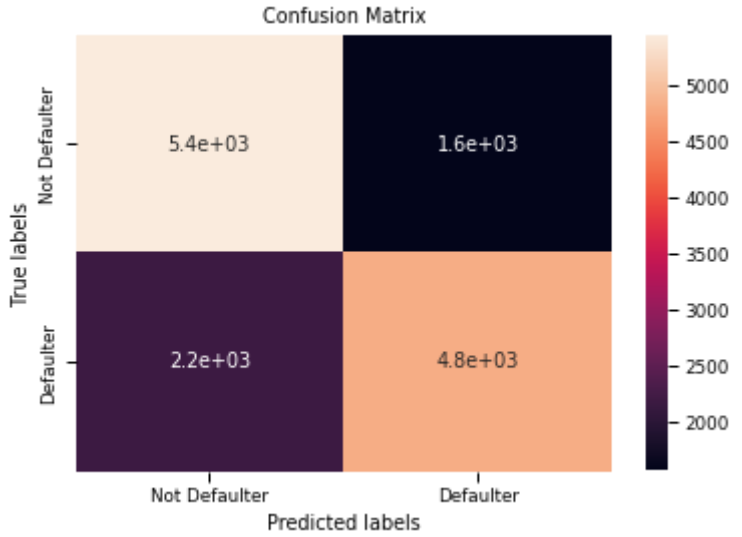
- The accuracy on test data is **0.750909**
- The precision on test data is **0.789361**
- The recall on test data is **0.684128**
- The f1 on test data is **0.732987**
- The roc_score on test data is **0.750876**



Decision Tree Classifier

Model Name	Accuracy Score	Precision Score	Recall Score	f1 score	roc_auc_score
Decision Tree Classifier	0.732934	0.755243	0.688838	0.720514	0.732912

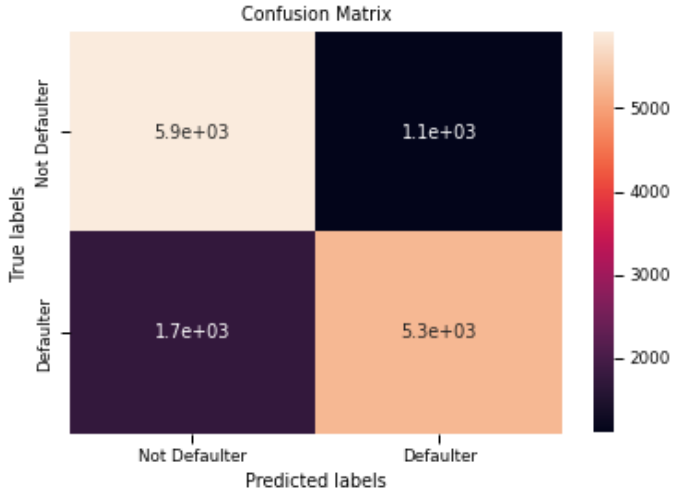
- The accuracy on test data is **0.732934**
- The precision on test data is **0.755243**
- The recall on test data is **0.688838**
- The f1 on test data is **0.720514**
- The roc_score on test data is **0.732912**



Random Forest Classifier

Model Name	Accuracy Score	Precision Score	Recall Score	f1 score	roc_auc_score
Random forest classifier	0.797204	0.827331	0.750928	0.78728	0.797181

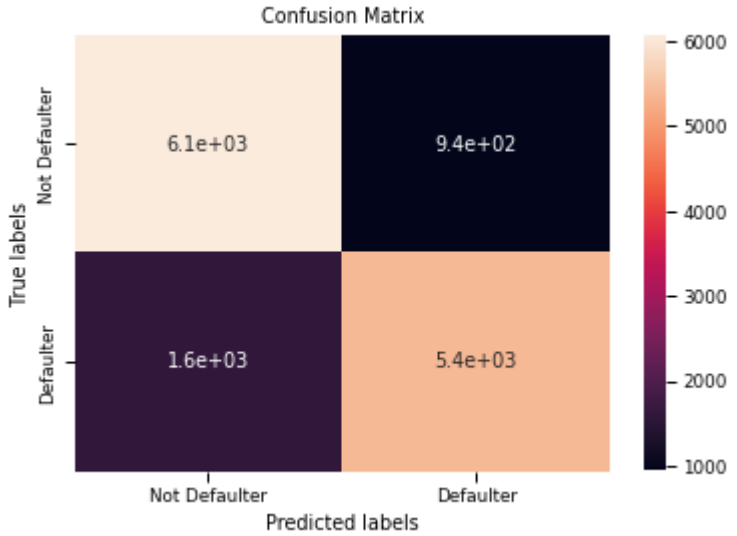
- The accuracy on test data is **0.797204**
- The precision on test data is **0.827331**
- The recall on test data is **0.750928**
- The f1 on test data is **0.78728**
- The roc_score on test data is **0.7971817**



XG Boost Classifier

Model Name	Accuracy Score	Precision Score	Recall Score	f1 score	roc_auc_score
XGB classifier	0.818247	0.851577	0.770625	0.809081	0.818223

- The accuracy on test data is **0.818247**
- The precision on test data is **0.851577**
- The recall on test data is **0.770625**
- The f1 on test data is **0.809081**
- The roc_score on test data is **0.818223**

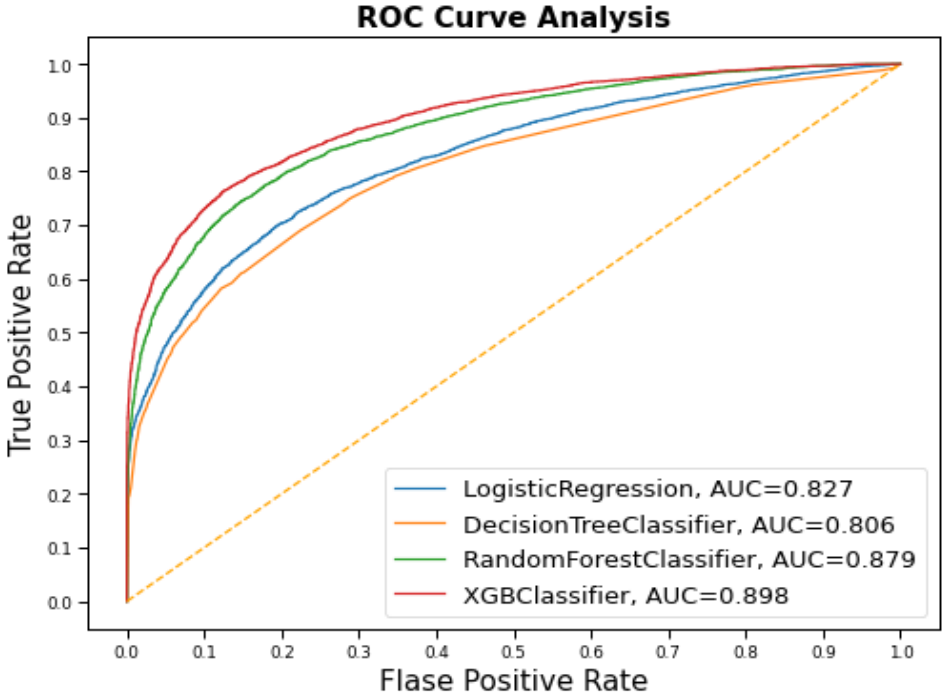


Comparison Of all model Evaluation Metrics

Model Name	Accuracy Score	Precision Score	Recall Score	f1 score	roc_auc_score
Logistic Regression	0.750909	0.789361	0.684128	0.732987	0.750876
Decision Tree Classifier	0.732934	0.755243	0.688838	0.720514	0.732912
Random forest classifier	0.797204	0.827331	0.750928	0.787280	0.797181
XGB classifier	0.818247	0.851577	0.770625	0.809081	0.818223

- XGBoost has the highest f1 score, precision, and recall after hyperparameter tuning.

ROC Curve Analysis Of All Models



- We deduced that the XGBoost and Random Forest Classifier algorithms provide better classification accuracy compared to traditional classification methods such as Logistic Regression and Decision tree classifier.

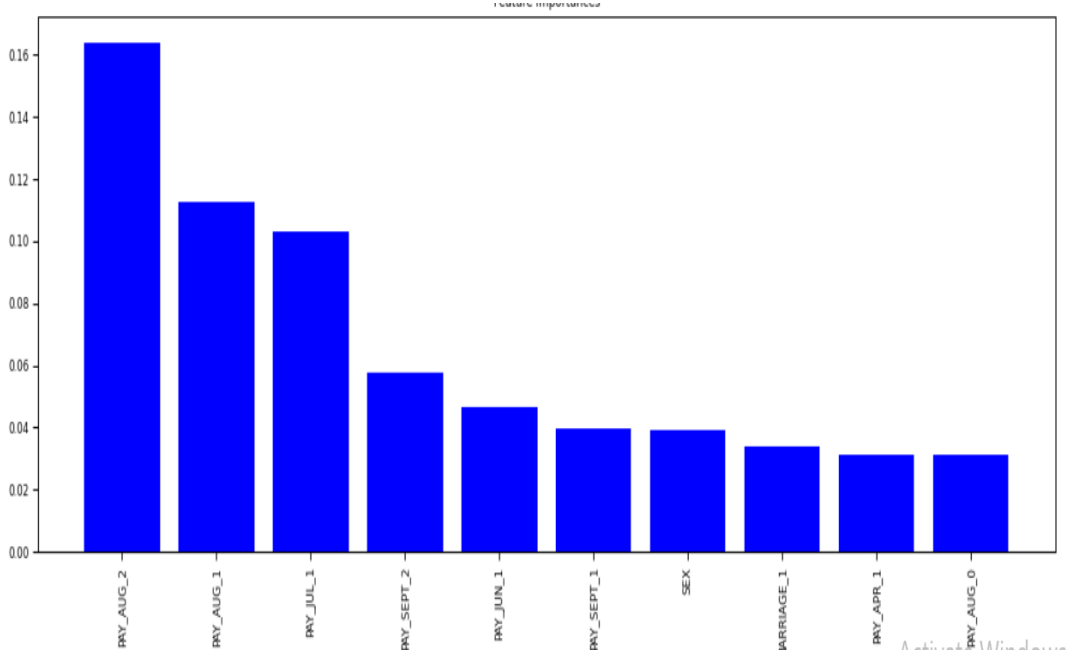
- Compared to the AUC of the Decision Tree Classifier at 0.806, the AUC improved to 0.898 for XGB Classifier.

Feature importance

XGBoost variable importance plot. The y-axis describes the percentage contribution of the predictor in the “real” model

The payment status on the previous months before default prediction month are the most important features.

The gender of the customers also play key importance in the prediction of the default



Conclusions And Observations

- Since the dataset is imbalanced, applying oversampling techniques such as smote helped in improving the model performance.
- Models without hyperparameter tuning showed higher accuracy for train data and its test accuracy is lower. It is suggesting overfitting of the model. Models with better results have a good balance between bias and variance.
- Hyperparameter Tuning is done on all models for higher performance and to prevent overfitting.
- Recall is the best evaluation metric for credit default prediction , because if the algorithm will not detect the defaulters, that will encounter more loss to the bank.

Contnd…

- As there is a trade-off between precision and recall, F1 score which is a harmonic mean of precision and recall is used as scoring for the models.
- XGBoost has the highest f1 score, precision, and recall after hyperparameter tuning. It is also the best model under ROC-AOC Curve.
- XGBoost is selected as the final model because it gives the best evaluation metric scores and has the highest KS statistic Score from the graph.
- Features showing recent months' payment status are the strongest default predictors. Sex and Marriage are the other important features.
- This Model can only serve as an aid in decision-making instead of replacing human decisions.
- Model can be improved with more data and computational resources.

Thank You !